North Carolina Education Research Data Center
Technical Report #5: Geocoding Addresses and Assigning MASTIDs.
June 17, 2019

Student address data provided by schools permit innovative analyses of the relationships between students' demographic characteristics, spatial environment, and academic achievement. This report documents the methods the North Carolina Education Research Data Center (NCERDC) employed to geocode the student addresses from the Transportation Information Management System (TIMS) databases and link to student MastID. Researchers can use this file to match student addresses to student-level data such as student test scores, demographics, and other datasets in the NCERDC archive. Using the method described below, the NCERDC was able to assign unique address IDs to many records.

Data

The TIMS data contains addresses for K-12 public school children from 1994 to present. Charter schools are excluded from these files. The number of years of available data varies by LEA, and the availability of data is not uniform over years, particularly in the 1990s.

Prior to 2007 four LEAs (from the 2005 PSU list) are not represented at all in the TIMS dataset: Cherokee (200), Forsyth (340), Mecklenburg (600) and Tyrell (890). Some other LEAs, which were consolidated prior to 2002, are also not represented in the TIMS data.

Cleaning the Data

From 1994 to 2009 the vast majority of the time required by this part of the project was spent in cleaning the TIMS data. A number of problems existed that had to be addressed before the data could be accurately geocoded and matched with the ABC data. Originally, NCERDC received over 400 files from TIMS, each with a name representing the LEA and year of the data. Of the 400+ LEA/year files received from TIMS, few were found to be unusable for one reason or another and discarded from our analysis. Duplicate files accounted for the majority of these cases. The good files were all read in and examined for accuracy. Problems that were found included:

Incomplete number of records
Files with garbage/nonsense records
Duplicate files
Incorrect LEA assignment
Incorrect year assignments
Incorrect grades
Differing record layout structure
Multiple LEAs (mostly cities within counties) embedded in one file with incorrect school codes
Sparse address data
Address problems: Missing zip codes, Incorrect zip codes, Ambiguous city codes, Missing city codes, Incorrect city codes, Non-standardized addresses

Solutions to these problems:

1. Each file's number of records was compared to the listed number of students found in the National Center for Education Statistics Public School Universe (PSU) data. Files whose record count differed significantly (either far more or fewer) from the PSU values were identified and fixed if possible. Files that could not be fixed were discarded.

2. A few files had lines of garbage text that had to be removed from the incoming files prior to processing.

3. Quite a number of files that were incorrectly labeled (e.g. ALAM98 and ALAM99) turned out to be exact duplicates. To determine which file was correctly named, we looked for matching students in the EOG files and based on year and grade we made a determination and renamed and/or discarded accordingly. We also compared all the files for one LEA with each other looking for internal consistency based on grades of students.

4. A few of the files were somehow assigned to the wrong LEA. These were quickly found since the addresses were not in the correct county. These were renamed correctly.

5. A big, yet subtle, problem consisted of files listed with the wrong year. Many of these were found from resolving the problems described above, but to be thorough we went through each file and compared a random sampling of students in the file to the ABC test data. Any anomalies found were either fixed or discarded.

6. In a few files, it appears that some of the records had incorrect grades listed. This is probably due to incomplete updating of the files by the school districts. We decided to keep these files and records since most of the information appears to be valid.

7. A number of the files had to be manually edited in order to bring the record layout structure into the same form as the other files.

8. Some of the files included multiple LEAs that needed to be identified so that the file could be parsed and the correct LEA and school codes assigned.

9. A good number of files had a large proportion of missing addresses or P.O. Boxes, which cannot be geocoded. If the files had fewer than 50% of usable addresses, they were not included in the analysis.

10. A significant amount of time was spent on pre-geocoding standardization of addresses.

Standardization Strategies

City name

The city variable in the TIMS data ranged from 1 to 4 characters with multiple codings possible for one place (e.g. B, BU, BUR, BURL, BTON all representing Burlington). A significant effort was made to create a variable with the entire city name using a two-step process. The first step involved finding and modifying a lookup database that contained records of each North Carolina and adjoining counties zip code (mapinfo data circa 2005) and its associated city and county. These measures were compared to the same variables in the address data, and if they matched then the complete city name was assigned. Unfortunately, a significant number of anomalies in the data made additional manual coding necessary. One reason was that zip code boundaries are not stable, so that an address in 1995 might not have the same zip code in 2004 (the date of the zip code lookup table). Also some of the city codes were not similar enough to the city name (e.g. 7lks for Seven Lakes). There were also some problems with multiple cities sharing the same zip code and having similar names.

Address

Standardizing the street address required a number of steps. First, all punctuation marks were removed from the street addresses. Then the address was broken into individual words which were each scanned for street types to standardize. The following standardizations were made:

Type	Standard	Scanned for
Avenue	AVE	AV, AVEE, AVEN, AVENUE, AVEUNE
Boulevard	BLVD	BLV, BOULEVARD
Circle	CIR	CI, CIRCLE
Court	CT	COUNT, COURT, CRT
Drive	DR	DRIVE
Highway	HWY	HY, HIGHWAY
Lane	LN	LANE
Parkway	PKY	PKW, PKWY, PARKWAY
Place	PL	PLACE
Road	RD	ROA, ROAD
Street	ST	STREET, STREETT, STT
Terrace	TER	TERR, TERRACE, TERRANCE
Trail	TRL	TR, TRAIL
Way	WAY	WY

Zip Code

Some records of the same address had more than one zip code, possibly resulting from data entry errors. To resolve these anomalies, we replaced all zip codes for a given address with the most common one. For example, if an address occurred three times, once with 27202 and twice with 27210, all three would be assigned 27210. Some zip codes were malformed in other ways as well. If they were blank, less than five digits, or the first digit was not 2 (i.e., it could not possibly be a North Carolina zip code), then the zip codes were replaced with good ones from the same address if they existed, otherwise they were left blank. We also included the zip codes from all the counties bordering North Carolina to account for the small minority of students who travel in from an adjoining state to attend North Carolina schools.

Geocoding

After as cleaning and standardizing addresses as possible, we created a subset file with only the unique addresses, assigned each a sequential id variable (addrid), and then appended this variable to the entire address file. Unique addresses were then geocoded using Centrus Desktop software. The geocoded results included all the original variables plus 11 new variables:

Block Group ID
Latitude
Longitude
New Address
New City
New Zip Code
New Zip+4
New County
New County FIPS code
New State

Match Code (A code indicating what type of match the geocoder used)
The match codes indicate what changes were made in the 'New' variables or if there was an error of some type. Examples of these codes were examined and based on those observations the geocoded addresses were divided into two groups – good and bad. The good addresses were then merged back with the entire address file by addrid.

Known geocoding problems:
Out-of-state addresses were not geocoded
The geocoder relies heavily on zip code and street address. If the zip code is wrong then incorrect changes may be made to the New Address and New City variables (along with the others to a lesser degree). We will continue to look for ways to minimize these types of errors.

In 2010 addresses were street geocoded with an address locator using Environmental Systems Research Institute's (Esri) StreetMap North America Smart Data Compression (SDC) dataset as reference. This reference dataset is based on 2005 Tele Atlas streets and enhanced by Esri.

NCERDC Amended Census Block ID

In an effort to further protect student's identities, an amended block id was created from the U.S. Census Block. In addresses that were geocoded prior to 2010 the census tracts were identified by a four-digit basic number and may have a two-digit numeric suffix; for example, 6059.02. The decimal point separating the four-digit basic tract number from the two-digit suffix is often shown in printed reports and on census maps (see: http://www.stssamples.com/census-tracts_block-group.asp). In cases containing the U.S. Census Block, the amended variable consists of the first 12 digits that can be linked to the census data (e.g., 371570605.003009 = amended block id 371570605003). In other cases the U.S. Census Block consists of only the census tract, such that the amended 11-digit variable indicates a specific tract but not a specific block group within that tract (e.g., 370010220.00 = amended block id 37001022000).
U.S. Census Block format before 2010

37001 0220.00 3007
| | |
| | |-Block Group
| |-Tract
|-County FIPS Code

In recent years the Census blocks are 15-digit codes containing the location's 2-digit state FIPS code, 3-digit county FIPS code, 6-digit census tract code, and 4-digit tabulation block code (see https://www.census.gov/geo/maps-data/data/baf_description.html). Census Blocks are contained in variables Block2000 and Block2010. The amended variable during these years contains the first 13 digits of the 15-digit block number. If fewer than 5 addresses can be linked to a 13-digit code, the amended variable was condensed to 11 digits, which equates to the census tract. This was determined to be the preferred method of addressing small localities after several other attempted solutions.

U.S. Census Block format after 2009

37 001 022000 3007
| | | |
| | | |-tabulation block code
| | |-Census tract code
| |-County FIPS Code
|-State FIPS Code

Imputations

From 1995 to 2009 simple imputations were done to increase the match rates with the ABC test data. This was done as follows. If a student was living at address A in year Y and living at the same address in year Y+2 and there is no data for year Y+1, then Y+1 was imputed to be address A as well. A new record was created, based on a copy of the previous year with the year variable being incremented by one. This same procedure was also used if 2 or 3 years separated the dates as well with multiple imputed records created. A flag variable marks these records (Addr_iflag = 1, 2 or 3 etc… indicating how many years of separation were imputed. Addr_iflag=0 means the record was not imputed). Currently the maximum value is 8 but this will get larger as additional years are added to the database. Most of the imputed data are created in the years in which a district has no data, i.e. they did not report a year or two. Additionally, years that do have district data may also contain some imputed records. Someone who does not want to use the imputed data can drop any record where addr_iflag>0 to limit the data to the non-imputed records.

The Matching Process

Matching records between the geocoded TIMS data and NCDPI student data involves various combinations of five variables - LEA, school, SSN, first name, and last name. Because SSN was not available (or was not valid) in many address records, matching had to rely on 4 or 3 of the variables. The order of matching was as follows:

Match 1 criteria = lea school ssn last first
Match 2 criteria = lea ssn last first
Match 3 criteria = lea last first spedis(ssn)
Match 4 criteria = lea ssn last spedis(first)
Match 5 criteria = lea ssn first spedis(last)
Match 6 criteria = ssn last first
Match 7 criteria = lea ssn last first (first-last transposed)
Match 8 criteria = lea ssn last first (first-middle)
Match 9 criteria = lea school last first

Spedis is a SAS procedure which tries to capture slight typos in a variable by looking for similar values.

Public Use Files

See AddressesLink.doc and AddressInfo.doc for more information about publicly-available datasets and variables.